Available online at www.sciencedirect.com
ScienceDirect
Energy Procedia 97 (2016) 447 – 453
Energy Procedia

European Geosciences Union General Assembly 2016, EGU
Division Energy, Resources & Environment, ERE

Data-driven surrogate model approach for improving the performance of reactive transport simulations

Janis Jatnieks[a,c,*], Marco De Lucia[b], Doris Dransch[a,c], Mike Sips[a,d]

[a] *GFZ German Research Centre for Geosciences, Section 1.5 Geoinformatics, Potsdam, Germany*
[b] *GFZ German Research Centre for Geosciences, Section 3.4 Fluid Systems Modeling, Potsdam, Germany*
[c] *Humboldt University of Berlin, Department of Geography, Berlin, Germany*
[d] *Humboldt University of Berlin, Department of Computer Science, Berlin, Germany*

Abstract

Geochemical simulation models are the computational bottleneck for coupled reactive transport simulations. We investigate the use of a data-driven surrogate model in place of a geochemical simulation model to speed up the run-times of reactive transport simulations. This is a challenge because the surrogate model needs to use results of its predictions as inputs at each subsequent simulation step. We test the suitability of surrogate model approach on a popular reactive transport benchmark problem, often used for evaluating simulation models. We show that the concept is feasible and can make the simulations many times faster, however several open areas for future work remain.

Peer-review under responsibility of the organizing committee of the General Assembly of the European Geosciences Union (EGU)

Keywords: Data-driven surrogate; reactive transport; coupled geochemical simulation; surrogate model

1. Introduction

Reactive transport models deal with simulation of geochemical reactions together with transport of fluids in the geological subsurface. These models have a wide range of important applications, such as the assessment of long-term outcome of underground gas storage [1], geothermal energy, and many others.

Reactive transport simulations are often implemented as a coupling of two distinct simulation models. One is responsible for the hydrodynamical processes - fluid flow and solute transport. Another is a geochemical simulation model responsible for geochemical reactions. Typically, geochemical simulations are the main computational bottleneck in coupled reactive transport simulations. This limitation allows only reactive transport simulations of coarse spatial resolution to be performed. In comparison, hydrodynamic simulations can have detailed geometries ranging into millions of elements that can be solved on single workstations [3]. Coupled reactive transport simulations of such detail are not currently feasible.

* Corresponding author. Tel.: +49-157-3245-1188.
E-mail address: jatnieks@janis.es

Peer-review under responsibility of the organizing committee of the General Assembly of the European Geosciences Union (EGU)
doi:10.1016/j.egypro.2016.10.047

The main reasons for the computational cost associated with geochemical simulations are the large number of parameter involved in each run and the large number of runs required at each simulation time step. Depending on the scenario and user interest, the number of input and output variables can range into hundreds [4]. In the framework of sequential coupling between hydrodynamics and chemistry, one geochemical simulation model is executed for every grid element at each time step of the coupled simulation. Executing the geochemical simulation that has such a large number of variables many times results in long simulation run times. For these reasons, the reduction of run times for geochemical simulations is of great interest for the scientific community.

2. Approach and challenge

One way to reduce the high computational cost of geochemical simulation is to replace the geochemical simulation model with a surrogate model. A surrogate model is a fast-running approximation that can be used as a replacement for the geochemical simulation model [5].

In this study, we focus on a data-driven surrogate approach. It is important to note that a data-driven approach is different from model reduction in mathematical modelling. Model reduction aims at preserving the underlying physical principles behind the simulation model while creating a simplified version of it. Instead, the data-driven approach considers input-output data produced from a coarse sampling of the parameter space of the simulation model, but without any physical considerations. As long as the surrogate model can predict the simulator output based on the input values, it may be possible to use the data-driven surrogate in place of the geochemical simulation model.

There are several reasons leading us to believe that only a necessary subset of geochemical models capabilities must be contained in a data-driven surrogate. First, geochemical simulations are often performed repeatedly for studying a specific site. This is computationally costly when using a geochemical simulation model. However, this site is associated with particular ranges of input and output data. When a reactive transport simulation for the site is performed with reduced spatial resolution, this provides input and output data samples from the geochemical simulation model. Second, the capabilities required from the surrogate model can be limited to a specific foreseen set of application scenarios. Finally, there is a margin for errors that may be acceptable if a perfect surrogate model is not obtainable. Geochemical models are affected by uncertainties concerning input parameters and the definition of a representative initial state for a particular site and scenario. Such uncertainties are inherent to the current state-of-the-art in geochemical simulations [6]. Because of this, the surrogate model may not always need to perfectly reproduce the simulator output. However, a surrogate model does have to be significantly faster than the "full physics" geochemistry simulator it is replacing.

To evaluate our approach in a reactive transport simulation scenario, we used a popular benchmark problem which has been defined to evaluate and compare different geochemical simulation models. Therefore, we believe it is also an appropriate test for a surrogate model.

There is a particular challenge associated with the use of a surrogate model instead of a "full physics" geochemical simulator for reactive transport. The surrogate model has to take its own prediction results, that are modified by the transport model, and use them as inputs at the next time step as shown in Figure 1. Most statistical models make some prediction errors. This can mean that with each subsequent time-step the changes in input parametrization can take the surrogate model outside the parameter space region for which it was trained. The large number of studies relying on surrogate models for various tasks in water resources research [5,7] do not use surrogate models in this recursive way. Our aim was to answer the question if it is feasible to use a surrogate model as a replacement for a coupled geochemical simulation model.

3. Related work

Surrogate modeling is a well established engineering approach for dealing with long-running simulation problems [5,8,9]. This can also be seen in the number of terms that are often used as synonymous for this approach. Among these are proxy models [10], emulators [11], meta-models [12], reduced order models [13], lower fidelity models [14] and response surface models [15]. Significant work has been invested in using surrogate models for speeding up computationally expensive simulations for water resources applications in general. The most comprehensive recent

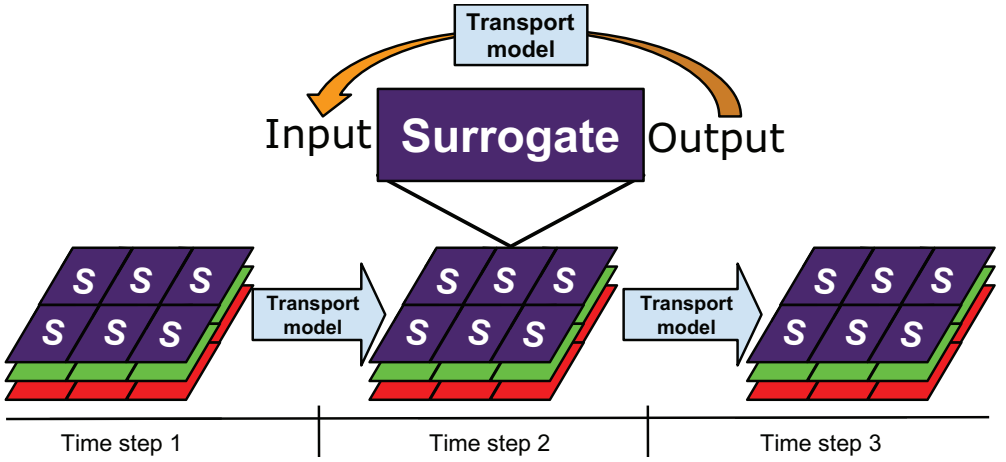

Fig. 1. The main challenge when coupling a surrogate model instead of a geochemical simulator for reactive transport simulations is the potential error propagation: the surrogate model has to re-use its results as inputs at the next time-step.

surveys of this are [5] and [7]. However, the vast majority of research surveyed in these articles does not concern reactive transport or geochemistry applications directly and is usually focused on hydrodynamic simulations alone.

When we restrict our attention to the use of surrogate models for reactive transport simulations or geochemistry, the number of recent contributions shrinks considerably. Genetic programming is used for learning the simulator response at specified sites in [16]. The approach in [13] employs non-linear regression with a quadratic polynomial for constructing the surrogate model from simulation results. In [17] the surrogate model approach is used for quantitative risk assessment purposes. In that paper the multivariate adaptive regression splines (MARS) algorithm is used as a surrogate method and its ability to reproduce the behaviour of the geochemical simulation model is investigated.

There are some common tasks for which the surrogates are often used. One very popular task is to assist an optimization or calibration problem for computationally expensive simulations [11,18–20]. In [12] the task for which the surrogate model is used is to enable global sensitivity analysis for long-running flow simulations, whereas in [10] the approach is called proxy modeling and is used for obtaining uncertainty analysis more efficiently. In such studies a surrogate model approach is used for speeding up the main task of the study, but not necessarily the simulation model itself. In contrast, the aim of our work is to create a surrogate model able to act as a replacement for a geochemical simulation model when coupled to hydrodynamics for reactive transport simulations.

4. Experimental set-up

In this study, we undertake a data-driven approach to the creation of surrogate models. Some studies focus on a particular method such as MARS [17] or genetic programming (GP) [16] and their capabilities to act as a surrogate model. Instead, we take a more method-agnostic approach based on the assumption that a good way to determine if a surrogate model is fit for its purpose is to try multiple methods and evaluate them. We trained 32 statistical and machine-learning methods available through the caret [21] and DiceEval [22] packages for open source language and programming environment R[1]. For each of these, we also tested seven pre-processing options available through the caret package. Overall this resulted in 1568 model selection screening runs, or 224 for each of the seven variables the surrogate models were trained for (Figure 2). This configuration is also sometimes known as a bucket of models. No additional attempt at combining or weighting the ensemble predictions is made. We also did not focus on exhaustively tuning or modifying the candidate methods.

Each of these combinations was trained using 7880 random samples amounting to 80% of the input-output data from the geochemical simulation model. To identify the best surrogate model candidates for each variable, we vali-

[1] www.r-project.org

dated the resulting surrogate model candidates using the remaining 20% of input-output data samples. During this, our system recorded several common error measures together with the execution time for each method and pre-processor combination. Among these error measures are absolute maximum error (AME), sum of absolute differences (SAD) and mean absolute scaled error (MASE) [23].

Once we have surrogate model candidates, we couple them into the reactive transport simulation instead of the geochemical simulation model. Our reactive transport case study consists of a common 1D benchmark, in which the injection of a reactive solution at one inlet triggers the dissolution of Calcite and the transient precipitation of Dolomite [24]. We implemented the benchmark relying on the PHREEQC geochemical simulation model through the Rphree interface [25,26]. It is important to note that the term "benchmark" is often used in two meanings. Here the meaning relates to a benchmark problem intended for testing how simulation models perform reactive transport problems. In section 5.2 the term "benchmark" relates to a timing benchmark that allows to compare how long the surrogate models and the simulation model perform a set of computations with the same parameter sets.

As a final step of surrogate model validation, we then replace the PHREEQC geochemical simulator in this reactive transport set-up with the surrogate model. We find that not every method showing good results according to the MASE error measure is able to replace the geochemical simulator in the reactive transport scenario. For this reason, we narrowed down the selection of methods to those that are able to complete the reactive transport simulation. This excludes methods that produce inconsistent output and those that produce only partial or no output during reactive transport simulation (not shown). We then repeated the reactive transport simulations using the surrogates that were able to complete this validation step.

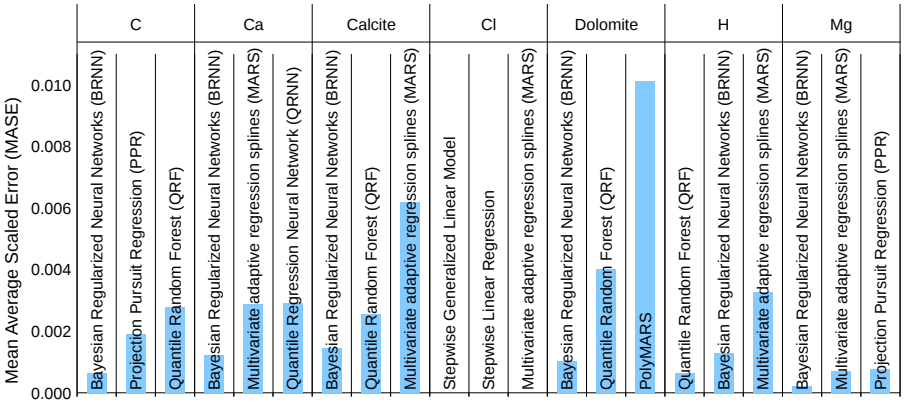

Fig. 2. We show the top three best performing methods, according to the MASE error measure, for each of the seven variables (top) that our surrogate models were aimed at predicting.

5. Results

5.1. Results from using surrogate models for coupled reactive transport

We used the MASE error measure for screening the best-performing prediction methods as candidates for predicting each variable supported by the surrogate model. In our experiments, this measure ranks the prediction methods similarly to other common measures, such as the residual mean squared error (RMSE) or the sum of squared errors (RSS). However, it has the advantage of allowing to compare prediction errors across variables of different value ranges [23] and never returning division by zero errors as other measures, such as Mean Absolute Percent Error (MAPE).

Our surrogate models are ensembles of prediction methods for seven variables involved in the reactive transport benchmark. Several methods perform well according to the MASE error estimator (Figure 2). To compare them in a reactive transport scenario, we tested the better performing methods by constructing surrogate models for replacing the simulation model in a reactive transport scenario. In Figure 3 we show some of the best results obtained using

different statistical methods. We find that Bayesian Regularised Neural Network (BRNN), shown in Figure 3b, is the best-performing surrogate model. Other methods and combinations of methods did not produce comparable results when used as a surrogate.

While our results show that a surrogate model can successfully replace a geochemical simulator in our simple reactive transport scenario, one important consideration is sample selection. Training statistical models with different random subsets of geochemical simulation model input-output data strongly affects the skill of the resulting surrogate model. Every surrogate model, from which the reactive transport results are shown in Figure 3, was trained on a different sample selection. Sample selection is also the only difference between the best- and worst-performing BRNN surrogates shown in Figures 3b and 3d. It can be seen in Figure 4b that the best-performing surrogate differs in that it is able to predict Dolomite better than others and the error values are more balanced for all variables. For our benchmark scenario we find that the BRNN and Quantile Regression Neural Network (QRNN) methods are able to provide a good overall trade-off between speed and accuracy (Figures 3 and 4a).

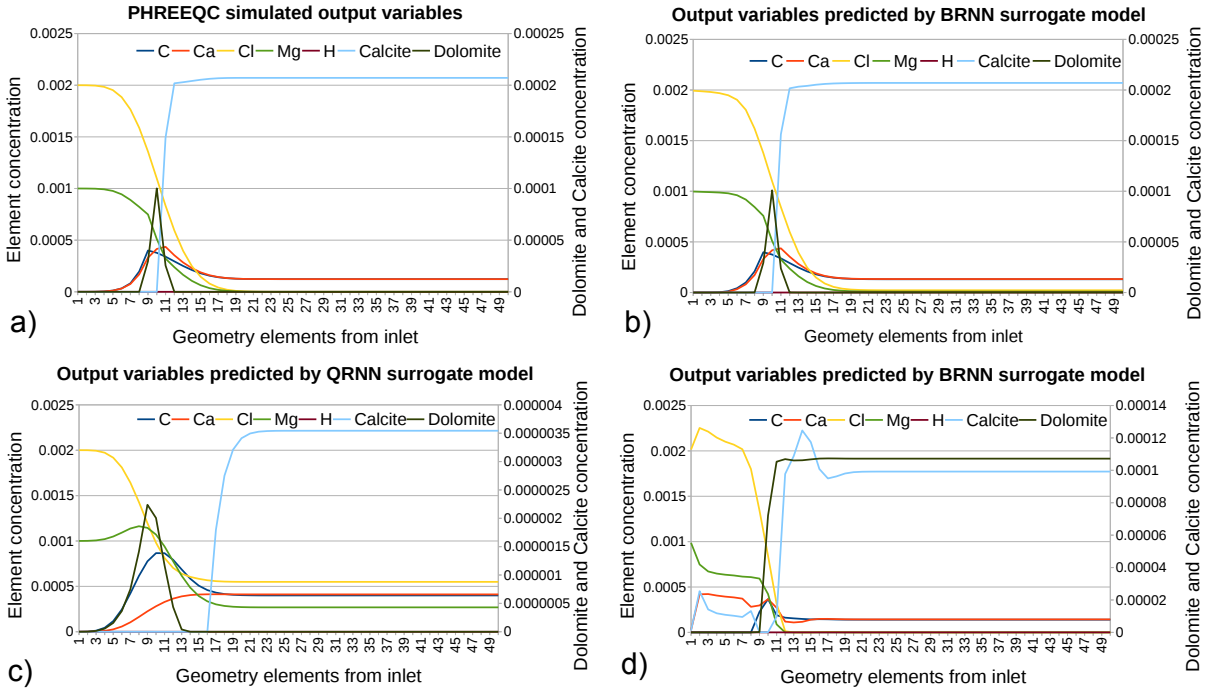

Fig. 3. Simulator (a) and surrogate (b) performing reactive transport, compared at time-step 100 of 197. An ensemble consisting of Bayesian Regularized Neural Network (BRNN) prediction models can act as a surrogate model for replacing the PHREEQC geochemical simulation model in this benchmark scenario. The surrogate model consisting of Quantile Regression Neural Network (QRNN) predictors (c) is shown as the second best. Another BRNN ensemble (d) is shown to illustrate how sampling affects the results.

5.2. Surrogate model speed benchmarks

It is worthwhile to look at the run time differences between the surrogate models and the full geochemical simulation model. We performed run time benchmarks for comparing the PHREEQC geochemical simulations with BRNN and QRNN surrogate model ensembles. We include these better performing surrogate model methods in the timing benchmark because a different scenario could have a different surrogate model as the best-suited for it. We measured the results using the R language profiler (Rprof) on a single core of an Intel i7-4702MQ CPU running at 2.2 GHz. The surrogate models and PHREEQC simulation model were run with the same parametrization, and the values showed in Figure 4 are obtained from averaging ten repeated measurements.

Since our aim was to test the feasibility of replacing the geochemical simulation model with a surrogate in order to decrease the computational time of reactive transport simulations, we relied on the R language due to its high

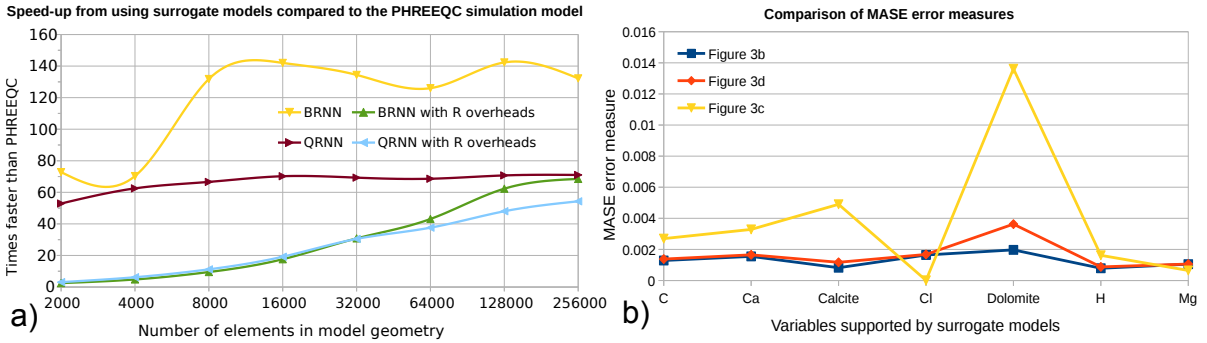

Fig. 4. The surrogate model that produces the best reactive transport results can be up to 140 times faster than the PHREEQC simulation model for this scenario (a). Comparison of MASE error measures (b) for the surrogate models shown in Figure 3.

abstraction level that is suited to prototyping. The series "with R overheads" in Figure 4 shows the speed-up achieved calling the surrogate model function in our R prototyping environment. The other data series show total time spent in the prediction function, which may be a better reflection of potential speed-up achievable with further optimization.

6. Conclusions

We used a simple reactive transport benchmark problem [24] as a simulation scenario for which we replaced the coupled geochemical simulation model with a surrogate model. To the best of our knowledge, this is a first time a data-driven surrogate model replaces a geochemical simulation model for a coupled reactive transport simulation. We showed that this is possible for our scenario and that a good agreement with the simulation model results can be obtained in this case. We also showed speed benchmarks to provide a basis for estimating potential speed-up that could be obtained for larger spatial geometries using the surrogate model approach described in this paper.

There are, however, several open problems that we learned in this process. First, it is important to note that the overall work-flow for obtaining a surrogate model depends on a number of considerations that can influence the skill of the surrogate model. This includes the combination of variables included and the prediction method chosen to serve as the surrogate model.

Second, the results strongly depend on the sample selection for training the surrogate model. This indicates the importance of an efficient simulation model parameter space sampling strategy for future work. Different prediction methods benefit from different selections of samples.

Third, as also stressed by other authors [17], the error measures used in the model selection step may not always directly map to an overall surrogate performance in the reactive transport application. The ultimate test for our surrogate model is to use it as a replacement for geochemical simulation model. Improving the model validation mechanism without the need to perform full reactive transport simulation could bring important future improvements for faster construction of geochemical surrogate models.

Finally, it is possible that each simulation site and scenario could be best approximated by a different surrogate model. Due to the large number of previously described considerations, in future work we aim to build an interactive visual analytics system that will allow this process to benefit from expert input and make it more systematic and streamlined towards the goal of data-driven surrogate model construction.

References

[1] De Lucia M, Kempka T, Kühn M. A coupling alternative to reactive transport simulations for long-term prediction of chemical reactions in heterogeneous CO_2 storage systems. *Geoscientific Model Development* 2015;8(2):279–294.

[2] Steefel C, Yabusaki S, Mayer U. Reactive transport benchmarks for subsurface environmental simulation, *Computational Geosciences* 2015:1–5.

[3] Virbulis J, Bethers U, Saks T, Sennikovs J, Timuhins A. Hydrogeological model of the Baltic Artesian Basin. *Hydrogeology Journal* 2013; 21 (4):845–862.

[4] Appelo CAJ, Postma D, Geochemistry, groundwater and pollution, CRC press, 2005.
[5] Asher M, Croke B, Jakeman A, Peeters L. A review of surrogate models and their application to groundwater modeling. *Water Resources Research* 2015;51 (8):5957–5973.
[6] Dethlefsen F, Haase C, Ebert M, Dahmke A. Uncertainties of geochemical modeling during CO_2 sequestration applying batch equilibrium calculations. *Environmental Earth Sciences* 2012;65 (4):1105–1117.
[7] Razavi S, Tolson BA, Burn DH. Review of surrogate modeling in water resources, *Water Resources Research* 2012; 48 (7).
[8] Müller J. Surrogate model algorithms for computationally expensive black-box global optimization problems, Ph.D. thesis, Julkaisu-Tampere University of Technology, 2012.
[9] Forrester A, Sobester A, Keane A. Engineering design via surrogate modelling: a practical guide, John Wiley & Sons, 2008.
[10] Josset L, Ginsbourger D, Lunati I. Functional error modeling for uncertainty quantification in hydrogeology, *Water Resources Research* 2015;51(2):1050-–1068.
[11] Sun Y, Tong C, Duan Q, Buscheck T, Blink J. Combining simulation and emulation for calibrating sequentially reactive transport systems. *Transport in porous media* 2012;92 (2):509–526.
[12] Rohmer J. Combining meta-modeling and categorical indicators for global sensitivity analysis of long-running flow simulators with spatially dependent inputs. *Computational Geosciences* 2014;18 (2):171–183.
[13] Bacon DH. Reduced-Order Model for the Geochemical Impacts of Carbon Dioxide, Brine and Trace Metal Leakage into an Unconfined, Oxidizing Carbonate Aquifer, Version 2.1, Pacific Northwest National Laboratory, 2013.
[14] Bianchi M, Zheng L, Birkholzer JT. Combining multiple lower-fidelity models for emulating complex model responses for CCS environmental risk assessment. *International Journal of Greenhouse Gas Control* 2016;46:248–258.
[15] Khuri AI, Mukhopadhyay S. Response surface methodology. *Wiley Interdisciplinary Reviews: Computational Statistics* 2010;2 (2):128–149.
[16] Esfahani HK, Datta B. Simulation of reactive geochemical transport processes in contamined aquifers using surrogate models. *International Journal of GEOMATE* 2015;8 (1).
[17] Keating EH, Harp DH, Dai Z, Pawar RJ. Reduced order models for assessing CO_2 impacts in shallow unconfined aquifers, *International Journal of Greenhouse Gas Control* 2016;46:187–196.
[18] Haftka RT, Villanueva D, Chaudhuri A. Parallel surrogate-assisted global optimization with expensive functions - a survey. *Structural and Multidisciplinary Optimization* 2016:1–11.
[19] Müller J, Shoemaker CA. Influence of ensemble surrogate models and sampling strategy on the solution quality of algorithms for computationally expensive black-box global optimization problems. *Journal of Global Optimization* 2014;60 (2):123–144.
[20] Razavi SS. Developing efficient strategies for automatic calibration of computationally intensive environmental models. Ph.D. thesis, University of Waterloo, 2013.
[21] Kuhn M. caret: Classification and Regression Training. R package version 6.0-70, 2016.
[22] Dupuy D, Helbert C, Franco J. DiceDesign and DiceEval: Two R packages for design and analysis of computer experiments. *Journal of Statistical Software* 2015;65 (11):1–38.
[23] Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. *International journal of forecasting* 2006;22 (4):679–688.
[24] Kolditz O, Görke UJ, Shao H, Wang W. Thermo-hydro-mechanical-chemical processes in porous media: benchmarks and examples. Springer Science & Business Media, Vol. 86, 2012.
[25] Parkhurst DL, Appelo CAJ. User's guide to PHREEQC (Version 2): A computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations. 1999.
[26] De Lucia M, Kühn M. Coupling R and PHREEQC: Efficient programming of geochemical models. *Energy Procedia* 2013;40:464–471.